Nuclear Instruments and Methods in Physics Research B 92 (1994) 188–193
North-Holland

Optimal estimation of uncertainty intervals for accelerator and decay counting [†]

L.A. Currie *

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Uncertainty concerning uncertainties reported by various investigators for counting experiments sometimes makes it difficult to assess the real meaning of experimental results, such as radiocarbon ages. Utilization of the maximum variance rule, which takes into account knowledge of counting error, has a major advantage in avoiding excessively small confidence intervals, but it leads to biased variance estimates and overly conservative confidence intervals. An assessment is given of the strengths and weaknesses of four alternatives, one of which "the variance weighted t" is asymptotically correct (negligible Poisson error) and is particularly attractive when Poisson error is dominant. Extension to the case where Poisson error is not the dominant, known random error component, shows that the methods presented can be generalized for non-counting experiments.

1. Introduction

Reporting of scientific results always entails two essential quantities: 1) an estimate of the underlying quantity, whether it be a length or a (radiocarbon) date, and 2) the uncertainty for that estimate. Uncertainty estimation can be an extremely difficult task, especially if reliable bounds for uncompensated systematic error are considered. Meaningful bounds for the random error component of the measurement process, however, can present a challenge; and one of the greatest problems is that different researchers commonly employ different prescriptions — sometimes unstated — for the random component of uncertainty. The purpose of this exposition is to examine the merits and pitfalls of some current and potential "random" uncertainty estimation practices for "counting" experiments. Issues involving systematic error will not be treated here.

For simple counting experiments, an ancient practice has been to assume Poisson counting statistics, or "shot noise", for the estimation of the standard deviation σ, and to construct confidence intervals with symmetric limits given by $x \pm U$, where the uncertainty $U = z\sigma$, x equals the measured value and z, the standard normal variate [#1]. The corresponding intervals will be optimal if there are no other random error components (H_0, "null hypothesis"), but overly optimistic (too short) if there are additional non-Poisson components (H_A, "alternative hypothesis"). For the vast majority of measurements in the physical sciences, limits are estimated using $U = ts$, where t is Student's-t and s, the square root of the estimated variance, s^2. Values of t are selected to achieve a confidence level $p\%$, based on ν degrees of freedom (df). For simple replication, where x and s are computed as the mean and standard error (standard uncertainty) of k observations, $\nu = k - 1$. These intervals will give the correct coverage (confidence level), but they are more disperse and generally wider than those based strictly on counting statistics — an unfortunate result if the null hypothesis (Poisson counting error only) happens to be true. For accelerator mass spectrometry (AMS), it has become common for some laboratories to use a maximum variance technique that takes the standard deviation as the larger of σ (Poisson) or s (replication) [2]. This provides important protection if the null hypothesis is not true, but the resulting confidence intervals [using max(σ,s)] will be overly conservative, especially if H_0 is true and degrees of freedom are few. It represents an exquisite example, however, of utilizing

[*] Tel. +1 301 975 3919, fax +1 301 216 1134, e-mail currie@enh.nist.gov.

[†] Contribution of the National Institute of Standards and Technology; not subject to copyright.

[#1] Use of "z" assumes that the expected number of counts is large enough that the Poisson distribution is approximately normal. Note that "U" is known as the "expanded uncertainty" [1]; it is equivalent to the half-width of the symmetric confidence interval.

Elsevier Science B.V.
SSDI 0168-583X(94)00124-E

knowledge of the measurement process for improved estimation.

In the following text we shall use a distribution sampling technique of $n = 1000$ x's and s's to present a visual comparison of the performance of the above three rules for (random) uncertainty estimation; also, we shall introduce a fourth rule that appears to offer even better performance under a variety of circumstances.

2. Objectives, issues, and alternative uncertainty rules

The objective is simple: to assess experimentally (computationally) the strengths and weaknesses of alternative rules for estimating (random) uncertainty for counting experiments, such as those involving radioactive decay or ion counting mass spectrometry. To accomplish this, the first step is to identify the performance issues, or "quality measures" to characterize uncertainty rule performance. Since all rules include some sort of variance (or standard deviation) estimate and some sort of "confidence" interval, we have selected the following:

- bias of the estimated variance [σ^2]
- "coverage", i.e. the actual confidence level achieved
- "exposure", i.e. the worst case normalized deviation from the truth: max(deviation/σ-estimate)
- relative size and dispersion of the confidence intervals

The first issue is what motivated this investigation, because the maximum variance method, popular among AMS laboratories, must give a positively biased variance estimate (in the long run) in the null case, because the distribution of estimates includes the long tail of s^2, but the short tail is cut off at σ_0^2 (the Poisson variance). Put differently, s will exceed σ_0 about half the time, yet no values smaller than σ_0 are allowed. This implies also that, for the null case, reported uncertainties will be overconservative. (In fact, they will be doubly overconservative, because of the use of Student's t.)

The uncertainty reporting rules examined are listed in Table 1. The fourth, "variance weighted-t" rule is included as an attempt to capture the scientific knowledge encompassed in Rule 3 (σ can never be smaller than the Poisson-σ), and at the same time temper its positive bias and overconservatism. This rule was inspired by a statistical method proposed for estimating confidence intervals involving multiple variance components [3], and it has been examined as a means for incorporating variance inequality constraints into detection and uncertainty estimation for physical and chemical counting experiments [4]. Certain limiting behaviors for the various rules are evident: 1) Assuming normally distributed random errors, Rule 2 is always unbiased, and coverage is as expected. 2) For the null case (Poisson error only), Rule 1 is unbiased and it gives the correct coverage; also it produces the shortest average confidence intervals. For all other cases, use of Rule 1 gives both variance estimates and confidence intervals that are too small. 3) When the Poisson error component is negligible, Rules 3 and 4 are asymptotically the same as Rule 2, i.e. they give unbiased variance estimates and correct coverage. 4) Differences among the rules are moot when degrees of freedom (df) are large, for then $t \to z$ and $s^2 \to \sigma^2$. 5) Rules 3 and 4 have the special advantage of limiting "exposure" (large deviation/σ-estimate) that is associated with the short tail of the s^2 distribution. Unfortunately, most counting experiments fall in a middle ground where non-Poisson error, which may be undetectable, is present; Rule 1 is then wrong, and Rules 3 and 4 are biased.

Table 1
Confidence interval rules [a] [when mixed counting and non-counting errors]

Rule 1:	Assume Poisson — σ only; CI $= x \pm z\sigma_0$
Rule 2:	Use replication — s only; CI $= x \pm ts$
Rule 3:	Use both, according to $s' = \max(\sigma_0, s)$; then CI $= x \pm z\sigma_0$ when $s \leqslant \sigma_0$, or $x \pm ts$ when $s > \sigma_0$
Rule 4:	Use a variance-weighted t' in connection with s' $t'_\mathrm{w} \equiv z(\sigma_0^2/s^2) + t(1 - \sigma_0^2/s^2)$

[The second term accounts for non-counting error variance, which can never be negative.]

[a] When counting error does not dominate, but an internal σ_i can be reliably computed, substitute σ_i for σ_0 in the above expressions.

2.1. Example

To illustrate the results of applying the four rules to real AMS data, we borrow data from ref. [5], where $k = 9$ replicate counts were made on a target prepared from ^{14}C reference material HOxII (SRM 4990C). The ratio of the average for HOxII to the HOxI standard (SRM 4990B) was 1.355 with a replication standard error of 0.00312, and a Poisson standard error of 0.00256 corresponding to a total of 280 thousand counts. The ratio (s/σ_0) is thus 1.22, so the four rules yield the following uncertainties:

$$\text{Rulc 1:} \perp (1.96)(0.00256) - 0.0050$$

$$\text{Rules 2 and 3:} \pm (2.306)(0.00312) = 0.0072$$

$$\text{Rule 4:} \pm (2.073)(0.00312) = 0.0065.$$

The critical value for (s/σ) for the detection of "excess variance" (σ_x^2, non-Poisson variance) for 8 df (5% significance level) is 1.39, so excess variance has not been detected. The detection limit for (σ_x/σ_0) is 2.16, while there is a 50% chance of detecting a ratio as small as 0.97 [6]. The several uncertainty estimates, differing by as much as 44%, typify the different levels of conservatism built into the rules and force the question as to which rule should be used. Had (s/σ_0) been less than 1, then Rules 1, 3, and 4 would have given the same result, and Rule 2 would have given a shorter interval. With smaller df, the differences are larger still. In fact, at higher levels of the replication tree (among targets, among laboratories, etc.) where one has a better chance of assessing *accuracy* in contrast to *precision*, small dfs are the rule. For this reason, the performance evaluation that follows was made with df = 4, i.e. 5 internal replicates.

3. Performance evaluation.

To set the stage for judging performance it is useful to view random deviations from the perspective of the different rules. Fig, 1a shows a 5% sample from 1000 normal random deviates (D) with zero mean and unit variance. Error bars represent $\pm\sigma_0$. Fig. 1b shows the same deviations, but with error bars $\pm s$ derived from a 5% sample of 1000 sets of replicates having 4 df each. Fig. 1c is similar to 1b but with the deviations ordered according to s. Fig. 1d is the same as 1c, except that s's smaller than σ_0 are replaced with that quantity (Rule 3). It is clear in Fig. 1a that Rule 1 yields an unbiased variance estimate and uniform width confidence intervals. Rule 2 applied to Fig. 1b or 1c also yields an unbiased variance estimate, but confidence interval widths are quite diverse. Unlike Rule 1 (Fig. 1a), Rule 2 shows occasional large "misses" (exposure) where Abs$(D/s) \gg 1$. For Rule 3 (Fig. 1d) the lack of balance between large and small s's leads to a positively biased variance estimate as well as overly conservative confidence intervals. On the other hand, exposure is controlled since s' cannot be smaller than σ_0. [Rule 4

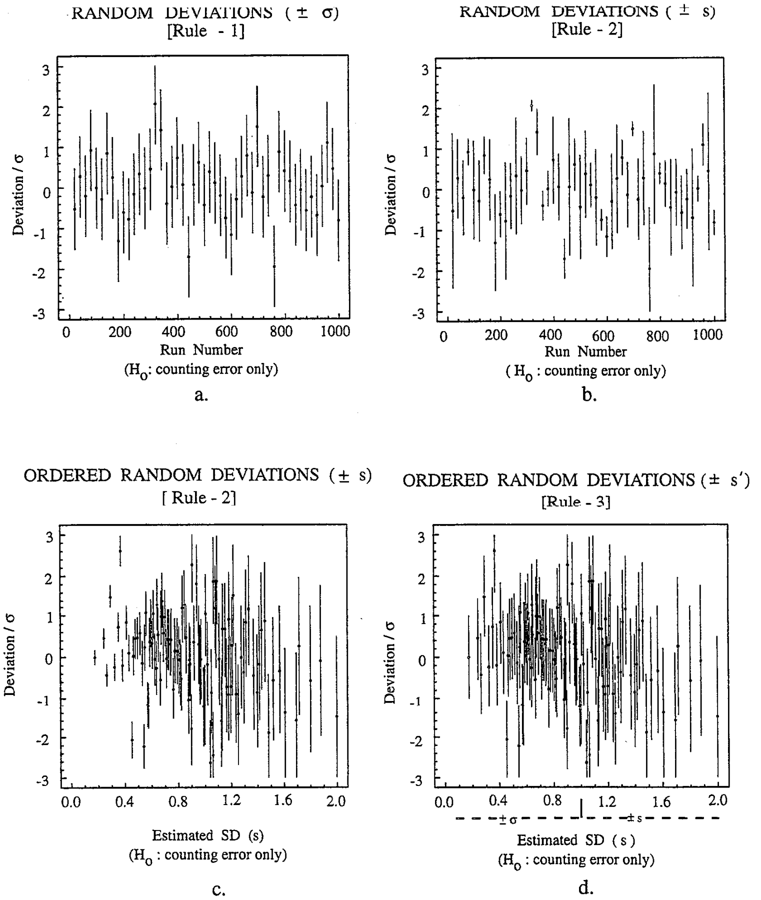

Fig. 1. Normal random deviates, 5% samples from 1000 simulations $N(0,1)$. (a) Random order, 1σ error bars; (b) random order, $1s$ error bars (4 df); (c) similar to (b), ordered by s; (d) with Rule 3 ($s \geqslant \sigma_0$) applied.

is similar to Rule 3 (Fig. 1d) except that the transition from z to t is more gradual.]

3.1. Variance bias

Bias is assessed by the direct computation of the variance from the 1000 replicates for the currently used rules and the null and alternative hypotheses, where H_A includes an extra variance component equal to the Poisson component, i.e. $\sigma_x^2 = \sigma_0^2$. For the 1000 simulations, the average ratios (±standard errors) of estimated to true variances are

	Ratio, Rule 1	Ratio, Rule 2	Ratio, Rule 3
H_0: count error	1.000	1.000±0.023	1.292±0.018 [#2]
H_A: added error	0.500	1.000±0.023	1.083±0.022

It is clear that Rule 1 performs badly (negative bias) when added random error is present, just as Rule 3 yields large positive bias when it is not. Rule 2 always yields unbiased variance estimates.

3.2. Coverage

Coverage refers to the probability that the confidence intervals (CI) will "cover" the true mean value. If the CI rule is overly optimistic, coverage will be smaller than the target confidence level; the converse is true if it is overly conservative. In this respect, the performance of the first three rules mirrors that obtained with the variance. The actual coverage for the set of 1000 CIs for Rule 2 was equal to the presumed level of 95% in both cases (H_0, H_A). For the null case Rule 1 also gave the correct coverage. Further insights into the validity of CIs and exposure are best illustrated by viewing cumulative distributions, as in Fig. 2. The only one of these rules that misses for H_0 is Rule 3. Fig. 2a shows the distribution of uncertainty normalized deviations for this case, and it is seen that the corresponding confidence intervals are too large, producing a coverage of 97% instead of the target 95%. (For H_A, Rule 3 gave a coverage of 94%.) Fig. 2b shows that Rule 1 errs seriously in the opposite direction (coverage = 84%) when additional random error is present (H_A). Fig. 2c is included to show that Rule 2 leads to correct coverage, but *exposure is severe*, with the extreme absolute deviations (D) greater than twice the 95% uncertainties (U). Recognizing that for 4 df, $U = 2.77\ s$, we find the most serious "misses" for the 1000 samples have $D = +7.22\ s$ and $-6.11\ s$. Direct integration of the t-distribution shows that *on the average* one should expect a miss $\pm 7.17\ s$ or more by chance for every 1000 samples! From the perspective of a large throughput of radiocarbon dates, say 2000–3000 per year, this would mean that with an $s = 100$ years, for example, by chance a few dates could be in error by as much as ±700 years. Rules 3 and 4 would prevent this *except* when the non-Poisson errors dominate, in which case Rules 3 and 4 are asymptotically the same as Rule 2.

[#2] An exact result for the Rule 3 (H_0) variance bias, courtesy of Mark Vangel and Keith Eberhardt (Statistical Engineering Division, NIST), for 4 df is $1 + 2e^{-2} \approx 1.271$.

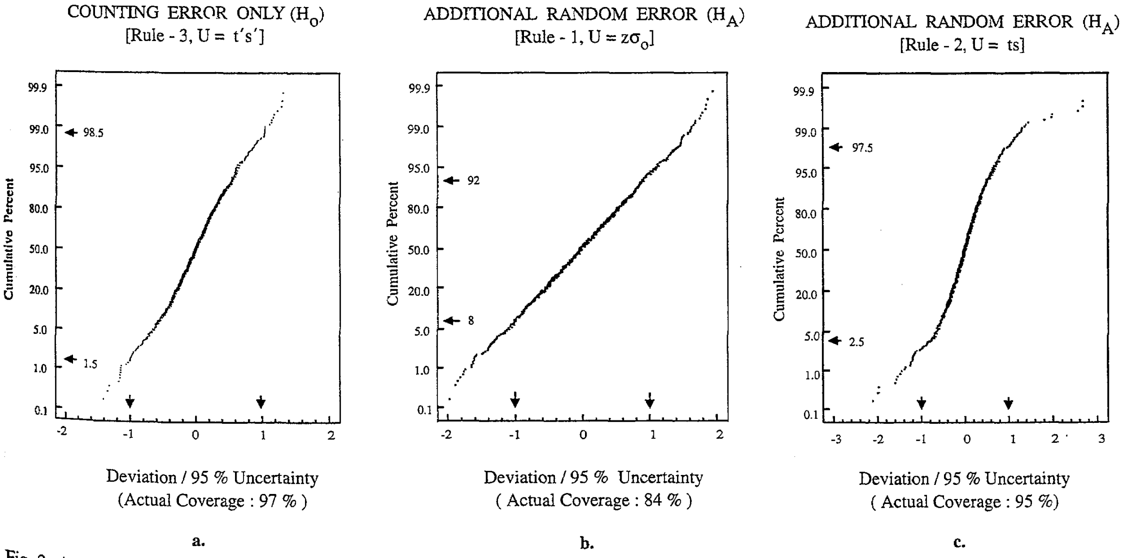

Fig. 2. Actual coverage of presumed 95% confidence intervals. (a) Null case ($\sigma = \sigma_0$), Rule 3; (b) alternative case ($\sigma > \sigma_0$), Rule 1; (c) alternative case, Rule 2.

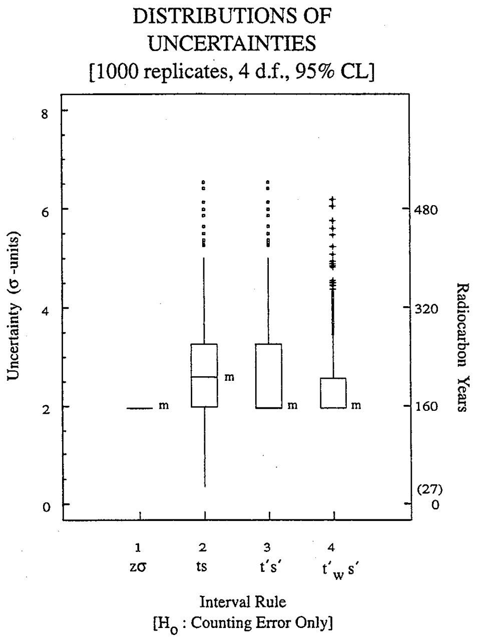

Fig. 3. Distributions of uncertainties, which represent confidence interval half-widths (null case; 4 df, 95% confidence level (CL); m = median). The scale at the left (95% uncertainty/σ) is perfectly general; the right ordinate is given for $\sigma = 80$ radiocarbon years.

3.3. Distributions of intervals

The final performance measure to be considered is the typical size and distribution of confidence intervals. This issue, which is most serious for the null case, is addressed graphically in Fig. 3. Metrics for the distributions are given in units of σ_0 on the left ordinate, and Radiocarbon years on the right, taking σ_0 as 80 years (1% σ for modern material). For Rule 1, there is no distribution; the 95% uncertainty is fixed at 1.96 σ_0, or 157 years. This is the best possible choice, if additional variance is *known* to be absent. Rule 2 is the worst choice; its 95% expanded uncertainties range from 27 to 522 years for the 1000 simulations. (Exact values computed from the F distribution are 33.4 and 477 years, respectively.) The central portion (interquartile range) extends from 159 to 261 years. Rule 3 eliminates the lower tail, such that the smallest interval, as well as the median, is the same as for Rule 1, 157 years. Rule 4, introduced in this study, is the best compromise, its median interval again equals that of Rule 1, but its central range is only half the size of that of Rule 3, with its upper quartile at 205 years.

4. Extension

When prior experience or theoretical knowledge of the measurement process indicates important non-Poisson random error components, then σ_0 should be replaced in Table 1 by a combined σ that includes all known random error components. This quantity, labeled "internal error" (σ_i) by some AMS laboratories [7], then serves as the base value or minimum possible σ. Rule 3, for example, then becomes: $s' = \max(\sigma_i, s)$. This formulation still applies in the limit of negligible Poisson error. Hence, using the σ_i extension, we find that the treatment investigated here is perfectly general, not limited to counting experiments.

5. Conclusion

Application of the alternative rules will lead to diverse and sometimes biased uncertainty estimates based on the *same* experimental data from decay and accelerator counting experiments. The most serious differences arise with small degrees of freedom and large confidence levels. For non-counting experiments lacking a prior or theoretical "internal" σ_i, there is little choice but to use the classic Rule 2, involving Student's-t. For AMS and decay counting, we conclude that

- Rule 1 is unquestionably the best if there is negligible non-counting (or other non-internal) error; otherwise it produces low estimates of imprecision and poor (confidence interval) coverage. The problem is *knowing* that Poisson (or internal) error dominates.
- Rule 2 is always unbiased, but it yields an extremely wide range of uncertainty estimates. For four degrees of freedom, the (95%) range of s's and hence uncertainties spans nearly a factor of five. High throughput laboratories must face the likelihood that deviations will exceed ±7s.
- Rule 3 gives maximum bias for the null hypothesis, since even then there is almost an even chance that s will exceed σ_0. The small s's that counter this excess (in Rule 2) are automatically excluded. The use of physical knowledge (that σ cannot be smaller than σ_0, or σ_i) is a major advantage, however, in suppressing the small s's and wide "misses" of Rule 2.
- As shown in the final, multi-box plot of uncertainties (Fig. 3), Rule 4 shows promise to reduce significantly the dispersion of uncertainty estimates compared to Rule 3, while still providing protection against unanticipated sources of random error.

In many respects, performance in the region where Poisson error dominates is most important, if there is concern about the efficient use of counting time, and once large external errors are brought under control

References

[1] B.N. Taylor and C.E. Kuyatt, Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, NIST Technical Note 1297 (1993).

[2] D.J. Donahue et al., Nucl. Instr. and Meth. B 29 (1987) 169.

[3] W.G. Cochran, Biometrics 20 (1964) 191.

[4] L.A. Currie (ed.), Detection in Analytical Chemistry Symp. Ser. 361 (American Chemical Soc., Washington, DC, 1988) chap. 1.

[5] D.B. Klinedinst, A.P. McNichol, L.A. Currie, R.J. Schneider, G.A. Klouda, K.F. von Reden, R.M. Verkouteren and G.A. Jones, these Proceedings (6th Int. Conf. on Accelerator Mass Spectrometry (AMS-6), Canberra–Sydney, Australia, 1993) Nucl. Instr. and Meth. B 92 (1994) 166.

[6] L.A. Currie, Nucl. Instr. and Meth 100 (1972) 387.

[7] R.P. Beukens, D.M. Gurfinckel and H.W. Lee, Radiocarbon 28 (1986) 229.